OLIVER EICHLER

Endlich Jarvis!

Deinen persönlichen KI-Assistenten selbst betreiben — lokal, souverän, mit Gedächtnis.

AGENTIC AI · SELF-HOSTED LLM

Java Forum Stuttgart 2026

45 MIN

Talk · DE

WER SPRICHT HIER

Oliver Eichler

Senior Cloud Manager · DigitalNativeAlliance · Stuttgart

ROLLE	Senior Cloud Manager
STACK	Kubernetes · AWS · Terraform
HEUTE	Self-Hosted LLM · Obsidian · Agents
DEMO	Zwei echte Bots · Live auf der Bühne

DevOps Engineer, Schwerpunkt Kubernetes, AWS, Systemarchitektur. Auf dem Weg vom Cloud Engineer zum GenAI Solution Engineer — und nehme euch heute mit in meinen Alltag.

AGENDA

45 Minuten.
Sieben Stationen.

01	Hook & Vision
02	Das Problem: KI kennt die Welt, aber nicht dich
03	Obsidian als geteiltes Gedächtnis
04	Technisches Fundament & die zwei Demo-Systeme
05	Der Stack: Vier Stufen zur Souveränität
06	Live: Demo & Das Duell
07	Closing: Erweiterbarkeit, Links & Danke

BLOCK 01

Hook & Vision.

DIE SUPERKRAFT

»Jarvis kannte alles über Tony Stark — seine Projekte, seine Geschichte, seinen Kontext.

Das war seine eigentliche Superkraft. Nicht die Stimme.«

DIE LÜCKE

ChatGPT kennt die Welt.

Aber dich?
Kein blasser Schimmer.

- Jede Session startet bei Null
- Eure Projekte: unbekannt
- Eure Entscheidungen: vergessen

Was ich heute zeige, läuft. Jetzt.

- Echte Hardware. Echte Agents.
- Kein Proof-of-Concept. Keine Demo-Umgebung.
- Ihr werdet es gleich selbst testen.

BLOCK 02

Das Problem.

DAS PROBLEM

LLMs sind stateless. Jede Session bei Null.

GUT BEKANNT — ABGEHAKT

- GitHub Copilot
- Claude Code
- Gemini Code Assist

→ Gut für Code. Im Editor. Fertig damit.

DIE ECHTE LÜCKE

- Dokumentation schreiben
- Entscheidungen nachvollziehen
- Planungen verfeinern
- Wissensarbeit jenseits des Editors

KERNTHESE

> »Ein Assistent ohne Gedächtnis ist teures Autocomplete.«

Und die Lösung ist verblüffend simpel.

BLOCK 03

Obsidian als Gedächtnis.

SO FUNKTIONIERT **MEIN** SECOND BRAIN

Lokal. Offen. Markdown. Deins.

MEIN PRINZIP

- PARA-Struktur + Zettelkasten
- Lokal auf deiner Hardware
- Markdown — gehört dir, läuft überall
- Direkte Basis für Agent-Zugriff

WARUM MARKDOWN?

- Kein proprietäres Format
- Lesbar für Menschen und Maschinen
- Keine Middleware, kein Protokoll
- Agent liest dieselben Files wie du

WO MEIN WEG LIEGT — ANDERE MACHEN DAS ANDERS

Klassischer Zettelkasten

Du formulierst und verknüpfst alles selbst.

Mein Weg →

Die KI macht die **Fleißarbeit — ich steuere**.

LLM-Wiki (Karpathy)

Die KI schreibt, verlinkt und pflegt alles.

DIE ENTSCHEIDENDE EINFACHHEIT

Gemeinsamer Wissenspool = .md-Dateien in einem Ordner.

DU	schreibst Notizen → `/vault/*.md`
AGENT	liest dieselben Files → denkt → schreibt neue Note zurück
SYMMETRIE	Gleicher Ordner. Gleiche Wahrheit. Kein Protokoll dazwischen.

KONKRETE USE CASES

Was das in der Praxis bedeutet.

01

Meeting vorbereiten

Aus eigenen Notizen, früheren Entscheidungen und offenen Punkten — automatisch zusammengefasst.

02

Entscheidungen finden

„Warum haben wir damals X entschieden?" — der Agent durchsucht den Vault und liefert Kontext.

03

Dieser Vortrag

Aus Stichpunkten zu Outline im Dialog. Die Notizen für diesen JFS-Vortrag kommen aus meinem Vault.

BLOCK 04

Technisches Fundament.

DER HARNESS-BEGRIFF

AGENT	Generiert Anfragen ans LLM
LLM	Verarbeitet → antwortet
TOOLS	`sitzen im Agent – nicht im LLM`
VAULT	Gemeinsames Gedächtnis

MEINE THESE

> **»Guter Harness + spezialisiertes Modell schlägt 0815-Agent mit Frontier-Model.«**

— mein Eindruck aus der Praxis. Die breite Validierung in der Branche läuft gerade erst an; ich teste sie selbst Projekt für Projekt.

DIE ZWEI DEMO-SYSTEME

Gleicher Zweck. Anderer Ansatz.

DEMO A — NOUS HERMES AGENT

HARDWARE	Raspberry Pi 5
HARNESS	Memory-Files — persistentes Gedächtnis
MODELL	Gemma 4 26B via OpenRouter
STÄRKE	Kontext über Sessions hinweg
ZUGANG	`@oeichler_pi_hermes_bot`
ALWAYS-ON	✓ 24/7

DEMO B — ANYTHINGLLM

HARDWARE	MacBook Pro M3 Pro 36 GB
HARNESS	RAG über Dokumente + Vault
MODELL	Gemma 4 12B QAT — lokal
STÄRKE	Präzisere Antworten durch Retrieval
ZUGANG	`Telegram → Bot B`
ALWAYS-ON	✗ nur wenn Mac läuft

REALITY CHECK

Warum nicht OpenClaw?

> ⚠ MÄRZ 2026 — SECURITY-KRISE
>
> 258.000 öffentlich exponierte Instanzen — 93% mit kritischen Auth-Bypass-Lücken. ClawHub Marketplace: malicious Skills von 341 auf 1.184 in zwei Wochen.

Wer in dieser Branche flexibel bleibt, gewinnt. Das ist auch eine Kompetenz, die ihr heute lernt.

LLM-INFERENZ — ZWEI PHASEN

Genug verstehen, um die richtigen Fragen zu stellen.

PHASE 1 — PREFILL

Prompt verarbeiten

- Alle Input-Tokens parallel
- Compute-intensiv
- **GPU-Kerne** entscheiden

PHASE 2 — DECODE

Antwort generieren

- Token für Token, sequentiell
- Liest bei jedem Token alle Gewichte
- **Memory Bandwidth** entscheidet

DER VERBORGENE SPEICHERFRESSER

KV-Cache: Je länger das Gespräch, desto mehr Speicher — für die Erinnerung, nicht für das Modell.

MODELL	GEWICHTE @ 4-BIT	KV-CACHE @ 64K	VRAM GESAMT	16 GB VRAM?
7B (Mistral, Llama 3.1)	~4 GB	~4 GB	~8 GB	⚠ Geht — kaum Headroom
12B (Gemma 4 12B)	~6 GB	~6–8 GB	~12–14 GB	⚠ Am absoluten Limit
26B (Gemma 4 26B)	~13 GB	~12–16 GB	~25–29 GB	✗ Unmöglich
70B (Llama 3.3)	~35 GB	~25–35 GB	~60–70 GB	✗ Unmöglich

MEMORY BANDWIDTH — WARUM APPLE HIER GEWINNT

Nicht Rechenleistung. Memory Bandwidth ist der Flaschenhals.

HARDWARE	BANDWIDTH	SPEICHER	URTEIL
Apple M3 Pro (dieser Mac)	150 GB/s	36 GB unified	⚠ 12B lokal flüssig — on par RTX 4090
Apple M5 Max	614 GB/s	48–128 GB unified	✓ **H100-Territorium — lokal, kein RZ**
NVIDIA RTX 5090	1.792 GB/s	32 GB GDDR7	⚠ Schnell — aber 70B unmöglich
NVIDIA RTX Spark (Herbst 2026)	300 GB/s	128 GB unified	⚠ M4 Max war schon schneller
NVIDIA H100 (RZ)	> 3.000 GB/s	80 GB HBM3	€2.093/Monat @ Scaleway — Daten verlassen das Haus

BUSINESS CASE AUF EINER ZAHL

MACBOOK PRO M5 MAX 128 GB

6.500€

EINMALIG · MODELL + DATEN BLEIBEN LOKAL

≠

NVIDIA H100 @ SCALEWAY (24/7)

2.093€

PRO MONAT · IHR BESITZT NICHTS · DATEN VERLASSEN DAS HAUS

Faustregel: 1B Parameter ≈ 2 GB @ 4-bit Quant — plus KV-Cache-Reserve einplanen

BLOCK 05

Vier Stufen zur Souveränität.

STUFE 1A — CLAUDE DESKTOP COWORK

Komfort.
Null Souveränität.

WAS ES BIETET

- Vault-Ordner einbinden, loslegen
- Kein Setup, kein API-Key
- Großartig zum Starten

⚠ 9. JUNI: FABLE 5 LIVE. 12. JUNI: US-EXPORTKONTROLL-DEKRET.

Drei Tage. Kein Vorlauf. Kein Einfluss. Alle Kunden — sofort, weltweit.

Das ist der Preis für null Souveränität.

STUFE 1B — CLAUDE + DISPATCH

Immer noch Claude.
Aber mobil.

HANDY	Fernbedienung — Telegram-Interface unterwegs
MAC	Maschine — Claude Desktop läuft, Vault bleibt lokal
DISPATCH	Persistenter Thread — verbindet Handy und Mac

Aber: immer noch Claude. Immer noch US-Cloud. **Immer noch abschaltbar.**

STUFE 2 — PI + NOUS HERMES + OPENROUTER

Kein Claude mehr.
Modellwahl bei mir.

```
Raspberry Pi 5 (Docker)
  ├── Nous Hermes Agent (WebSearch · WebExtract)
  ├── OpenRouter (LLM-Proxy: Gemma 4 26B / Mistral / DeepSeek)
  └── Tailscale (sicherer Kanal von überall)

Vault ← GitHub / iCloud / Samba  ·  Zugang: Telegram → @oeichler_pi_hermes_bot
```

Noch Abhängigkeit: OpenRouter als Proxy. Aber: **welches Modell, welcher Anbieter — das entscheide ich.**

STUFE 3 — MACBOOK + LM STUDIO + ANYTHINGLLM

Kein API-Key. Kein Anbieter. Volle Souveränität.

```
AnythingLLM (Agent + RAG + UI)
  ↓
LM Studio @ localhost:1234
(Metal GPU · Gemma 4 12B QAT)
  ↓ tool calls
DuckDuckGo · Web Fetch · Vault
```

TOKEN PANIC — JUNI 2026

Copilot: Flatrate → Token-Abrechnung. Bis zu **€1.800/Monat**. Uber: KI-Budget 2026 in vier Monaten verbrannt.

Self-hosted: kein Token-Zähler. Keine Rechnung.

ZUSAMMENFASSUNG

Souveränität ist kein Schalter — eine Reise.

STUFE	STACK	SOUVERÄNITÄT	AUFWAND
1a	Claude Desktop Cowork	✗ Null	zwei Stunden
1b	Claude + Dispatch	✗ Gering	ein Abend
2	Pi + Hermes + OpenRouter	◐ Modellwahl	ein Wochenende
3	MacBook + LM Studio + AnythingLLM	✓ Vollständig	ein Nachmittag

BLOCK 06 · LIVE

Live-Proof.
Es läuft.
Wirklich.

SCREEN WECHSELT: SLIDES → DEMO

DEMO-SETUP

Zwei Bots.
Zwei Antworten. Live.

→ Jetzt wechsle ich den Screen von den Slides auf die Live-Demo.

BEAMER-SETUP

- Beide Telegram-Chats nebeneinander
- Bot A: `@oeichler_pi_hermes_bot`
- Bot B: AnythingLLM via Telegram
- Beide online. Beide bereit.

WARM-UP-FRAGE

"Was war die Kernthese meines JFS-Vortrags?"

Memory vs. RAG — woher kam die Antwort? Wer war schneller?

DAS DUELL — IHR FRAGT, DIE AGENTS ANTWORTEN

Stellt eine Frage. Ich schicke sie an beide.

- Wissensfrage (Web-Recherche) → hier glänzt RAG
- Persönliche Vault-Frage → hier glänzt Memory
- Kombination → wer verknüpft besser?

LIVE · BLOCK 06

Live.

DEMO A

@oeichler_pi_hermes_bot

DEMO B

AnythingLLM · Bot B

Zum Abschluss.

ERWEITERBARKEIT

Text ist mein Fundament — und erst der Anfang.

HEUTE — TEXT & MARKDOWN

Mein Second Brain ist textlastig. Was Markdown lesen kann, wird Teil des Stacks:

- E-Mails: zusammenfassen, Draft-Replies, IMAP
- Planung: Kalender & Blog-Artikel aus Notizen
- Workflow: GitHub Actions, Helm Charts, Skripte

MORGEN — VISUELLER KONTEXT

Moderne LLMs verarbeiten Bilder hervorragend — rein visueller Kontext wird nutzbar:

- Screenshots statt nur Text
- Diagramme, Whiteboards, Skizzen
- UI-Zustände direkt erfassen

> ⚠ ABER: MEIN KLARES NO-GO
>
> Automatische Desktop-Screenshots, wie manche Agents sie schon machen. Zu oft ist Sensibles sichtbar, das nichts im Token-Stream verloren hat — solange der nicht **rein lokal** läuft.

EINSTIEG — HEUTE ABEND

Die wichtigsten Links.

GEDÄCHTNIS	obsidian.md
AGENT + RAG	anythingllm.com
INFERENCE	lmstudio.ai
LLM-PROXY	openrouter.ai
ARTIKEL	medium.com/@oeichler

DAS WAR'S — FAST

Danke!

Sprecht mich gerne noch hier auf dem JFS an — oder vernetzt euch via LinkedIn.

EUER EINSTIEG — HEUTE ABEND

- Obsidian + Claude Cowork: **zwei Stunden** (wenn der Vault schon steht)
- AnythingLLM + LM Studio: **ein Nachmittag**
- Nous Hermes auf Pi: **ein Wochenende**

→ VERNETZEN AUF LINKEDIN

Oliver Eichler

linkedin.com/in/oeichler

medium.com/@oeichler